一种自动分类的网页搜索排序算法 *

刘铭瑀，刘学亮，胡　骏

(合肥工业大学 计算机与信息学院, 合肥 230009)

摘　要：针对传统网页排序算法 Okapi BM25 通常会出现网页与查询关键词领域无关的领域漂移现象，以及改进算法需要人工建立领域向量的问题，提出了一种基于 BM25 和 Softmax 回归分类模型的网页搜索排序算法。该方法首先对网页文本进行数据预处理并利用词袋模型进行网页文本的向量表示，之后通过少量的网页数据来训练 Softmax 回归分类模型，来预测测试网页数据的类别分数，并与 BM25 信息检索的分数结合在一起，得到最终的网页排序结果。实验结果显示该检索算法无须人工建立领域向量，即可达到很好的网页排序结果。

关键词：领域向量；BM25；Softmax 回归分类；网页排序

中图分类号：TP391.1　　doi: 10.3969/j.issn.1001-3695.2017.07.0700

Web page search ranking algorithm using automatic classification

Liu Mingyu, Liu Xueliang, Hu Jun

(*School of Computer & Information*, *Hefei University of Technology*, *Hefei* 230009, *China*)

Abstract: In the traditional Web page ranking algorithm Okapi BM25, there exists a problem that the retrieval results are independent to the domain keywords, and the improved algorithm needs to build the domain vector manually. To address this issue, we propose a web page ranking algorithm based on BM25 and softmax regression classification model. In this method, we first encode the web page text with the bag-of-words model, and then train the softmax regression classification model by a small amount of web data to predict the category scores of the test web data. Finally we combine the category scores and the BM25 information retrieval scores to get the final ranking of web page results. Experiment results show that our method can meet the user's information need better without even manually creating the domain vector.

Key Words: domain vector; BM25; softmax regression classification; Web page ranking

0　引言

随着互联网爆炸式的发展，Web 信息在每个人的生活中变得越来越重要，然而当面对大量的信息时，用户从中找到有用的信息就严重依赖于搜索引擎的功能了，所以网页排序算法一直是搜索引擎的研究热点。但是，用户检索的关键字往往是很简短且不精确的[1]，导致了搜索引擎中高排名的网页可能与用户搜索意图并不相关[2]。例如，用户搜索关键词“微博”，有的可能是想搜索到“微博”的登录界面，有的则是想得到“微博”这家公司的新闻、股票等相关信息。互联网上的内容涵盖了多个主题，在现实生活中，人们想要得到的网页返回信息往往是某一主题内的，这就是所谓的领域问题。实际影响搜索排序的因子有很多，信息检索是最主要的因素之一。多年以来，许多研究学者在信息检索领域做了大量的工作，提出了布尔模型[3]、向量空间模型[4]和概率模型[5]等许多有代表性的信息检索模型，布尔模型和向量空间模型都将文档表示词条视为相互独立的项，忽略了表示词条间的关联性，概率模型则考虑了词条、文档间的内在联系，利用词条之间以及词条与文档间的概率相依性进行信息的检索，而 Okapi BM25 算法[6]作为概率模型的典型排序算法，已经在搜索引擎的网页排序[7, 8]、自然语言处理的文本加权[9, 10]等领域得到广泛使用。

近年来，大多数基于 BM25 的相关性排序算法主要利用词频，例如 TF、IDF 信息等来计算查询与网页之间的相关性，TF 指的是词条在文档中出现的次数，IDF 通过统计包含词条的文档数量来衡量词条的重要性。Büttcher 等人[11]在 BM25 的基础上加入了临近信息模型，该模型计算了词条在文档中的距离信息来改善 BM25 的评分。Roi-Blanco 等人[12]同样改进了 BM25 算法并应用到网页检索当中，该模型通过考虑词条的不同来源来计算文档的词条的重要性，并通过在“虚拟区域”上定义运算符来计算词条与文档的相似度。上述方法都将 BM25 算法或其改进算法应用到了网页的检索排序之中，并取得了不错的效果，但是这些方法并没有有效的解决领域漂移。针对这个问题，

基金项目：国家自然科学基金资助项目（61472116，61502139）；安徽省自然科学基金资助项目（1608085MF128）

作者简介：刘铭瑀（1992-），男，黑龙江齐齐哈尔人，硕士，主要研究方向为数据挖掘与人工智能（qiqihaerlmy@gmail.com）；刘学亮（1981-），男，河北石家庄人，副教授，主要研究方向为多媒体信息处理；胡骏（1990-），男，安徽合肥人，博士，主要研究方向为多媒体与机器学习.

文献[13]提出了基于领域模型的网页排序算法(topic sensitive re-ranking, TSRR)，该算法设计了一种独立于网页排序的模型，模型能够选取领域关键词组成的向量来表示领域，然后建立网页信息模型，在用户检索过程中结合领域向量模型和网页信息模型对网页搜索结果进行重排序。该算法效果的好坏取决于领域关键词选择的多少以及关键词建立的准确性。然而领域关键词的建立比较费时费力，而且严重依赖专业知识和直觉。

本文针对上述方法存在的问题，提出了一种自动分类的网页搜索排序算法。所提算法与前述算法不同点在于无须人工建立领域向量，而是采用分类器自动获得网页的类别概率。在信息检索方面，本文使用 BM25 算法来计算检索关键词与网页的相关性。领域方面，用少量的网页训练 Softmax 回归分类模型，得到网页数据的领域概率分数。将 BM25 分数和领域概率分数线性相加，对网页进行排序。该方法无须经验和人工技巧，即可以达到一个很好的网页排序效果。

1　本文方法

本文方法如图 1 所示，首先用爬虫程序爬取网页文本，之后进行分词、去停用词和词袋模型向量化等预处理，形成了本实验的数据集。接着训练 Softmax 回归分类模型得到每个网页的类别概率。根据用户提供的搜索关键词，通过 BM25 算法检索相关网页，将网页的 BM25 分数和 Softmax 的类别分数进行融合，得到最终的网页分数，以此排序来将相关网页返回给用户。

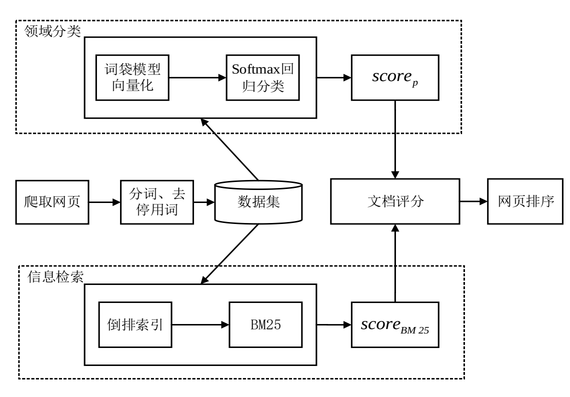


图 1　本文算法流程图

1.1　数据预处理

首先，通过爬虫程序对不同领域的网站进行爬取，并根据网页的标签提取出文本内容，然后采用结巴分词中的精确模式进行中文分词，并去掉停用词。

在数据预处理后得到了已分词后的网页文本数据集，对通过中文分词切分成词语的文本进行计算，得到每个词语在每个网页文本中的权重，此处用到了自然语言处理中的词袋模型(bag-of-words)。如果某个词语在网页文本中出现 n 次的话，则在网页文本向量中对应的权重值为 n，否则为 0。网页文本向量的大小为$m * |v|$，其中 m 为网页文本的个数，$|v|$表示每一个文本向量的长度，具体大小为词典中词的个数。

1.2　Okapi BM25 算法

Okapi BM25 是一个经典概率模型计算公式，它根据给定搜索查询词与匹配文档的相关性对文档进行排名。给定一个索引向量 Q，包含关键词$q_1, q_2, \ldots, q_n$，对于一个文档 D，BM25 的分数公式为：

$$score_{BM25} = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D)(k+1)}{f(q_i, D) + k\left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \tag{1}$$

其中：$f(q_i, D)$是关键词q_i在文档 D 中的词频，$|D|$是文档 D 的长度，$avgdl$ 是平均文档长度，k 和b是两个可调节的参数，一个决定了词频的比重，一个决定了文档长度的比重。实验验证[14]，通常将 k 设置为[1.2, 2.0]，本文为 1.2，b 设置为 0.75。$IDF(q_i)$是检索词q_i的逆文档频率，公式为

$$\mathrm{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{2}$$

其中：N 是采集的文档总数，$n(q_i)$为包含关键词的文档个数。观察该公式可以看出，如果一个词q_j在一半以上的文档里都出现，那么$IDF(q_j)$为负值，所以本文在数据预处理阶段，就把停用词都去掉。

为了提高查询效率，减少响应时间，本文采用倒排索引机制(inverted index)[15]，在全局搜索下，倒排索引可以建立并存储词条(term)与文档(doc)之间的关系映射。通过倒排索引，可以根据词条快速获取包含这个词条的文档列表。在经过数据预处理过程之后，每篇文档都转换成一个词条列表<term, doc>，对所有文档按照其中出现的词条来建立倒排索引，索引中包括一部词典，和一个全体倒排记录表。如图 2 所示。

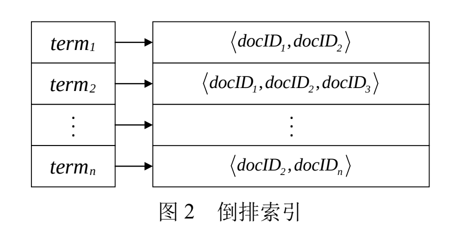


图 2　倒排索引

1.3　网页文本分类

网页信息是包含多个领域数据的，因此，网页分类是一个多分类问题。本文采用 Softmax 回归分类模型对网页进行分类。Softmax 回归分类模型有很多优点，例如它是直接对分类可能性进行建模，无须事先假设数据分布，这样就避免了假身份不准确所带来的问题，而且它不仅可以预测出类别，还可以得到近似的概率预测，这就对本文的任务很有帮助。该模型数学定义如下：

$$p\left(y^i = j \middle| x^i\right) = \frac{e^{W_j x^i}}{\sum_{l=1}^{k} e^{W_l x^i}} \tag{3}$$

其中：W 是模型参数，x^i为第 i 个网页文本向量，$p\left(y^i = j \middle| x^i\right)$表示 Softmax 回归将 x 分类为类别 j 的概率。对于给定的输入实例 x，按照上述的条件概率分布求得各类别的概率，取概率最大者，选为其类别。

对于爬取到的五个类别的网页，本文首先将全部的网页加上类别标签，为了近似达到线上模型的效果，随机抽取 1500 篇的数据来训练网络，余下的 13500 篇数据作为测试数据来测试分类准确率，测试数据与训练数据不交叉。经过调参，当准确率达到最优时，固定好网络的参数，将 13500 篇测试数据再次输入到网络中，得到 Softmax 回归分类模型预测的类别概率输出 $\boldsymbol{p}$。

领域分类方面，采用余弦相似度来计算 Softmax 回归分类模型的类别概率输出 $\boldsymbol{p}$ 与类别向量 $\boldsymbol{l}$ 的相似度，来得到类别分数。Softmax 回归分类模型的类别概率输出 $\boldsymbol{p}$ 与类别向量 $\boldsymbol{l}$ 可以表示为如下形式：

$$\boldsymbol{p} = (p_1, p_2, \cdots, p_n)$$

$$\boldsymbol{l} = (l_1, l_2, \cdots, l_n)$$

其中：n 为类别个数，类别向量 $\boldsymbol{l}$ 中只有一个元素的值 $l_i(i = 1,2, \cdots, n)$为 1，其余元素为 0。余弦相似度的计算公式如下：

$$\text{score}_p = \frac{\sum_{i=1}^{n} p_i * l_i}{\sqrt{\sum_{i=1}^{n} p_i^2} * \sqrt{\sum_{i=1}^{n} l_i^2}} \tag{4}$$

1.4　算法整体描述

自动分类网页排序算法的具体步骤如下：

a)利用式(1)(2)计算用户搜索关键字和每个网页的相似度，得到分数$score_{BM25}$。

b）首先用 Softmax 回归分类模型对网页进行概率预测，其次利用式(4)计算 Softmax 模型的类别概率输出 $\boldsymbol{p}$ 与类别向量 $\boldsymbol{l}$ 的余弦相似度，得到类别分数$score_p$。

c)对两个分数进行加权求和，得到网页排序的最终分数公式：

$$score = \alpha \cdot score_{BM25} + \beta \cdot score_p \tag{5}$$

其中:$\alpha + \beta = 1$，α，$\beta \in [0,1]$，根据 score 对网页进行排序，得到最终的网页排序结果。自动分类网页排序算法的具体步骤如下：

Input：query, page, l

Output：result

```
for each page in pagelist{
    score_BM25=BM25(query, page)
    p=SOFTMAX(page)
    score_p=cosine-relative(l, p)
    score=α·score_BM25 + β·score_p
    result.add(score, page)
    result.sorted(score)
}
return result
```

2　实验结果及分析

2.1　实验设置

本文实验所使用的机器配置为 Intel(R) Xeon(R) CPU E5-2620@2.10GHz，RAM 64 GB，Ubuntu 14.04 操作系统，算法采用 Python 2.7 实现。

2.2　实验检索关键词及语料

本文使用网络爬虫从腾讯、新浪、IT 时代网等常用网站爬取了 IT、创业、学术、时政、娱乐等五大类新闻语料。总共 15000 篇，每一类 3000 篇。检索关键词挑选自 2016 互联网各领域热词，每个领域选取三个检索词。实验数据具体选择如表 1 所示。

表 1　语料来源及检索关键词

领域	语料来源	检索关键词
IT	IT 时代网	虚拟现实、比特币、无人机
创业	IT 时代网、科技讯	云计算、直播、共享单车
学术	合肥工业大学新闻网、中国科学技术大学新闻网	基因、生物、人工智能
时政	腾讯、新浪、搜狐	一带一路、互联网+、文化自信
娱乐	八卦来了	王宝强、杨洋、粉丝

2.3　评价标准

为了验证本文排序方法的有效性，实验采用了如下两个评价指标：

a)用户满意度。用户对排名前十的每个网页进行打分，采用五分制，分数为 1~5，分别表示很不满意、不满意、一般、满意、十分满意。最后计算前十网页的打分均值。公式如下：

$$Satisfication = \frac{1}{n}\sum_{i=1}^{n} S_i \tag{6}$$

其中：n 是用户数量，S_i表示用户对排名前十的网页打分的均值。

b)Precision at K。P@K 是信息检索领域一个最直观的指标，它反映了检索回的前 K 个结果中被认为是相关的文档的比例。所以该指标衡量的是用户对整体检索结果的满意度。根据文献[16]的评价标准，本文选择在用户满意度指标中网页评分 3 分以上的，即满意和十分满意的网页作为相关的网页检索结果。P@K 的公式如下：

$$\text{P@}K = \frac{K_s}{K} \tag{7}$$

其中，K_S指的是前 K 个查询结果中相关网页的个数。对于检索系统而言，用户想要的结果排名越靠前，那么这个检索系统越是成功的，为了将排序位置信息也加入到评测指标中，本文分别选择 P@2、P@4、P@6、P@8、P@10 来判断检索结果的好坏。

2.4　参数调优

式(5)决定网页排序的最终得分，对于两个参数α，β需要根据实验结果来权衡哪一部分的比重更大，实验的方法是找五个志愿者，每个领域选取出一个关键词，利用用户满意度公式，对不同的参数计算出来的排序结果进行评分，并确定参数的最

优值。每个关键词进行九次实验，参数分别从第一次的$\langle 0.1, 0.9\rangle$到第九次的$\langle 0.9, 0.1\rangle$。图 4 是实验结果。

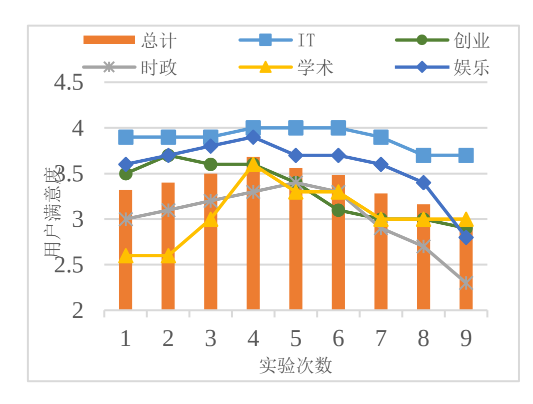


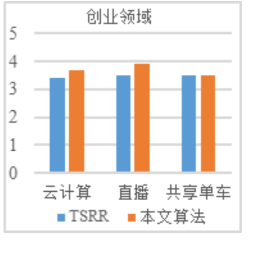
图 4　参数调优结果

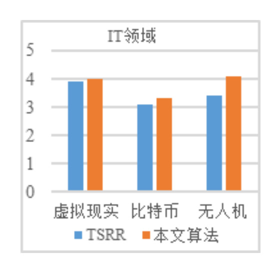
根据实验结果，IT、学术、娱乐三个领域在$\langle 0.4, 0.6\rangle$时用户满意度分数最高，同时创业、时政也达到了第二高评分，五类的平均分也是在$\langle 0.4, 0.6\rangle$时达到最高，因此本文选择$\alpha = 0.4$，$\beta = 0.6$作为最优参数值。

Softmax 回归分类器的迭代次数为 50 次，步长设置为10^{-4}，当分类准确率达到最高约为 94.6%时，获得分类器的最优参数，将测试数据输入网络中计算得到类别概率。

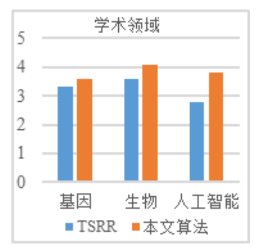
2.5　对比实验结果

本文选择的对比方法同样是为了解决领域漂移问题的 TSRR 算法，由于数据集不同，该算法同样按照本文参数调优的方法确定参数最优值为$\alpha = 0.3$，$\beta = 0.7$。在事先不告知两种排序结果分别属于哪种算法的前提下，找五名志愿者，对所给各领域检索关键词，按照对某个关键词在对应领域想要得到的检索结果对每个关键词所检索回的网页结果进行用户满意度打分。图 5 是实验结果对比。由图 5 可知，一些关键词本身包含一定领域信息，例如：IT 领域的"虚拟现实"、"比特币"，创业领域的"云计算""直播""共享单车"、学术领域的"基因""生物"，时政领域的"一带一路""文化自信"，娱乐领域的"王宝强""杨洋""粉丝"等。这些关键词的结果 TSRR 的用户满意度平均分为 3.53，本文算法获得了 3.87，比 TSRR 高 9.6%。但是像"无人机"既可以是 IT 领域，也可以是创业领域。"人工智能"既可以是 IT 领域，也可以是创业领域或者学术领域。"互联网+"也是如此。这些关键词在检索的时候就可能存在严重的领域漂移现象，这三个词 TSRR 平均分为 2.83，本文算法平均分为 3.73，比 TSRR 提高了 31.8%。有效的解决了领域漂移的问题。









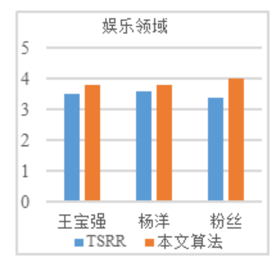


图 5　各领域用户满意度实验结果

表 2　P@K 实验结果

算法		TSRR					本文算法				
	关键词	P@	P@4	P@6	P@8	P@10	P@2	P@4	P@6	P@8	P@10
IT	虚拟现实	1	1	0.83	0.88	0.8	1	1	**1**	**1**	0.8
	比特币	1	0.5	0.5	0.5	0.4	1	**0.75**	0.5	0.5	**0.5**
	无人机	0.5	0.5	0.33	0.38	0.4	**1**	**1**	**0.83**	**0.88**	**0.8**
创业	云计算	0	0.25	0.17	0.25	0.4	**1**	**1**	**0.83**	**0.88**	**0.7**
	直播	0.5	0.5	0.33	0.38	0.5	**1**	**1**	**0.83**	**0.88**	**0.9**
	共享单车	0	0.25	0.33	0.25	0.4	**0.5**	**0.5**	0.33	0.25	0.4
学术	基因	0.5	0.75	**0.7**	0.63	0.6	**1**	0.75	0.5	0.63	**0.7**
	生物	1	1	0.83	0.88	0.7	1	1	0.83	0.88	**0.9**
	人工智能	0	0.25	0.17	0.13	0.2	**1**	**0.75**	**0.83**	**0.75**	**0.8**
时政	一带一路	1	1	0.83	0.75	0.7	1	1	0.83	**0.88**	**0.9**
	互联网+	0	0.25	0.17	0.13	0.1	**1**	**0.75**	**0.67**	**0.63**	**0.6**
	文化自信	0.5	0.25	0.5	0.63	0.7	**1**	**1**	**1**	**1**	**1**
娱乐	王宝强	1	0.75	0.67	0.63	0.6	1	**1**	**0.83**	**0.75**	**0.7**
	杨洋	1	1	0.83	0.75	0.6	1	1	0.83	0.75	0.6
	粉丝	0	0.25	0.33	0.5	0.6	**1**	**1**	**1**	**0.88**	**0.9**

本文选择 P@k 来衡量算法对检索结果位置的好坏。结果如表 2 所示，P@2 提升了 81.3%，P@4 提升了 58.8%，P@6 提升了 54.8%，P@8 提升了 50.5%，P@10 提升了 45.5%。本文算法的 P@2 等指标有较大提高，确保了用户想要得到的结果

返回在较靠前的位置。综上所述，Softmax 回归分类模型与 BM25 结合的网页排序算法即有效的解决了网页排序中领域漂移问题，也使得相关网页的排序更加靠前。

3 结束语

本文提出了一种结合 BM25 和 Softmax 回归分类模型的网页排序算法，算法采用少量的网页数据训练分类器，获得类别分数，与BM25检索分数相结合，得到网页排序的最终分数。该方法无须人工建立领域向量，有效的解决了领域漂移的问题，同时能够保证了相关的网页排名更加靠前。在后续的研究中，将用户的历史搜索、行为倾向[17~19]等加入到该算法当中，使得网页排序算法得到更好的效果。

参考文献：

[1] Fonseca B M, Golgher P B, Moura E S D, et al. Using association rules to discover search engines related queries [C]// Proc of LA-WEB. 2003: 66-71.

[2] Zhuang Ziming, Cucerzan S. Re-ranking search results using query logs [C]// Proc of the 15th ACM International Conference on Information and Knowledge Management. 2006: 860-861.

[3] Cooper W S. Getting beyond boole [J]. Information Processing and Management, 1998, 24 (3): 243-248.

[4] Salton G, Yang C S, Yu C T. A theory of term importance in automatic text analysis [J]. Journal of the American Society for Information Science and Technology, 2010, 26 (1): 33-44.

[5] Robertson, Stephen E, Jones S, et al. Relevance weighting of search terms [J]. Journal of the Association for Information Science and Technology, 2014, 27 (3): 129-146.

[6] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond [J]. Foundations and Trends® in Information Retrieval, 2009, 3 (4): 333-389.

[7] Niu Jianwei, Zhao Qingjuan, Wang Lei, et al. OnSeS: a novel online short text summarization based on bm25 and neural network [C]// Proc of IEEE Global Communications Conference. 2016: 1-6.

[8] Li Ying, Sha Fei, Wang Shujuan, et al. The improvement of page sorting algorithm for music users in Nutch [C]// Proc of the 15th IEEE//ACIS International Conference on Computer and Information Science. 2016: 1-4.

[9] Bestgen Y. Improving the character n-gram model for the DSL task with BM25 weighting and less frequently used feature sets [C]// Proc of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects. 2017: 115-123

[10] Kazemian S, Zhao Shunan, Penn G. Evaluating sentiment analysis in the context of securities trading [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2016: 2094–2103.

[11] Büttcher S, Clarke C L, Lushman B. Term proximity scoring for Ad hoc retrieval on very large text collections [C]// Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2006: 621-622.

[12] Blanco R, Boldi P. Extending BM25 with multiple query operators [C]// Proc of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2012: 921-930.

[13] 潘澄，吴共庆，李磊，等. 基于领域模型的网页搜索排序算法 [J]. 计算机系统应用, 2015, 24 (11): 107-114.

[14] Jones K S, Walker S, Robertson S E. A probabilistic model of information retrieval: development and comparative experiments [J]. Information processing and management, 2000, 36 (6): 809-840.

[15] Manning C D, Raghavan P, Schütze H. An Introduction to Information Retrieval [M]. Cambridge: Cambridge University Press, 2008: 1-18.

[16] Agichtein E, Brill E, Dumais S. Improving Web search ranking by incorporating user behavior information [C]// Proc of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2006: 19-26.

[17] Wu Yao, DuBois C, Zheng A X, et al. Collaborative denoising auto-encoders for top-n recommender systems [C]// Proc of the 9th ACM International Conference on Web Search and Data Mining. 2016: 153-162.

[18] Wu Chaoyuan, Ahmed A, Beutel A, et al. Recurrent recommender networks [C]// Proc of the 10th ACM International Conference on Web Search and Data Mining. 2017: 495-503.

[19] Zhuang Fuzhen, Luo Dan, Yuan N J, et al. Representation Learning with Pair-wise Constraints for Collaborative Ranking [C]// Proc of the 10th ACM International Conference on Web Search and Data Mining. 2017: 567-575.